



# A collective agenda on AI for Earth sciences

[Manil Maskey, Ph.D.](#)

# What I hope to do today

Data systems perspective on AI for Earth science

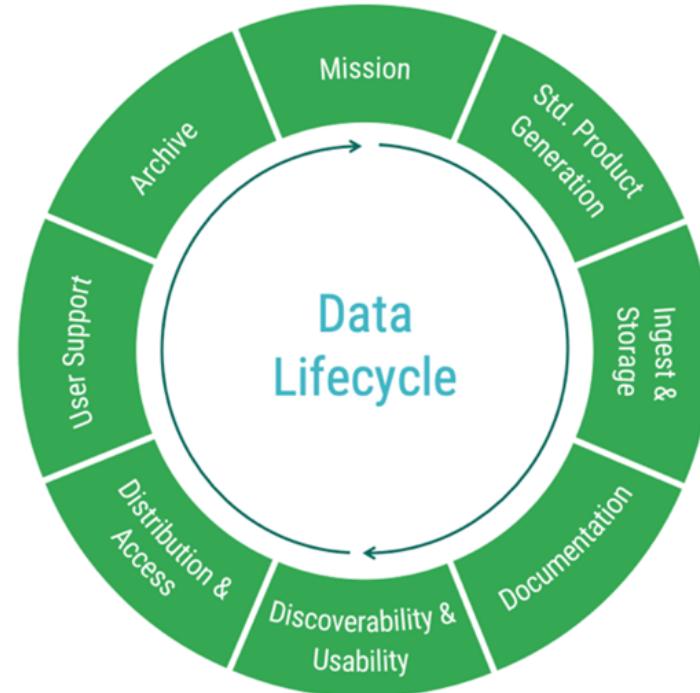
- AI for core data systems services
  - Search
  - Knowledge discovery
- Enabling AI to advance Earth science
  - Data (labeled training data) is the proprietary differentiator
  - Transitioning AI models to Production
  - Citizen science

○ ○ ○

# NASA's Earth Science Data Systems Program

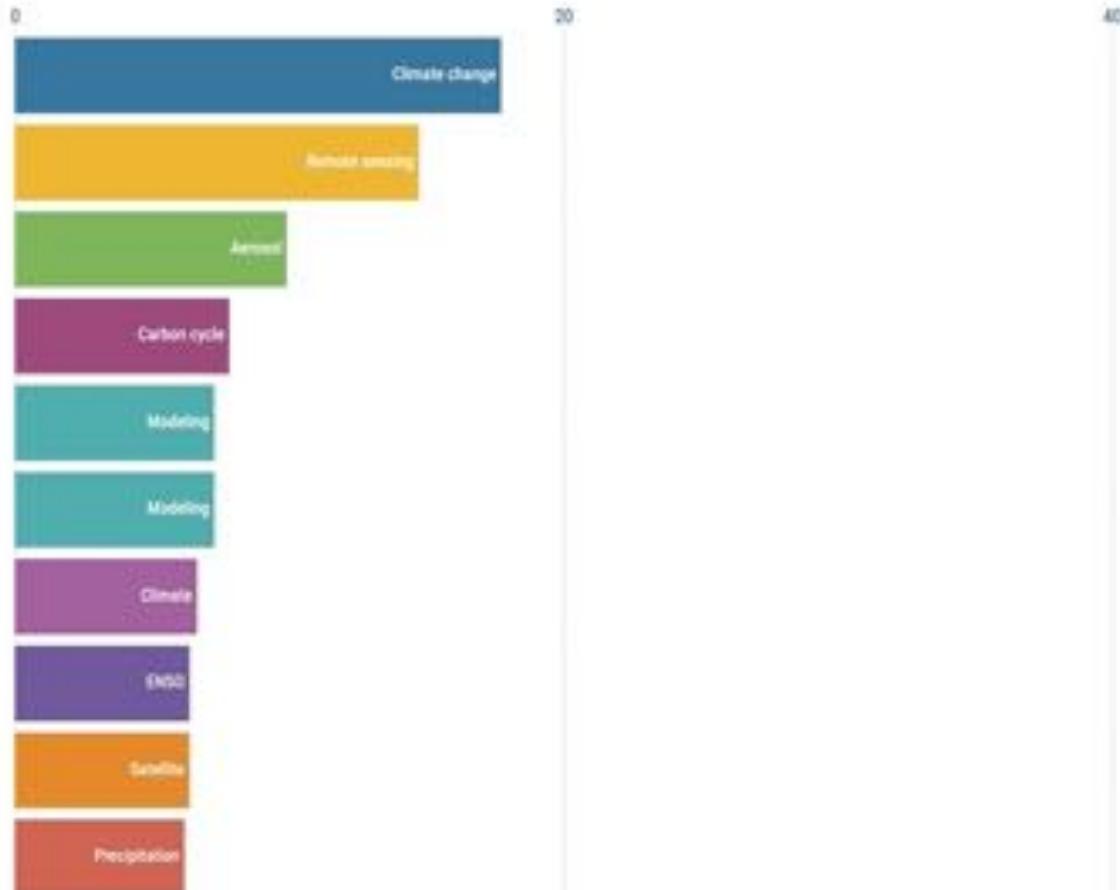
Single largest repository of Earth Science Data

Manages NASA's Earth science data through the entire data life cycle



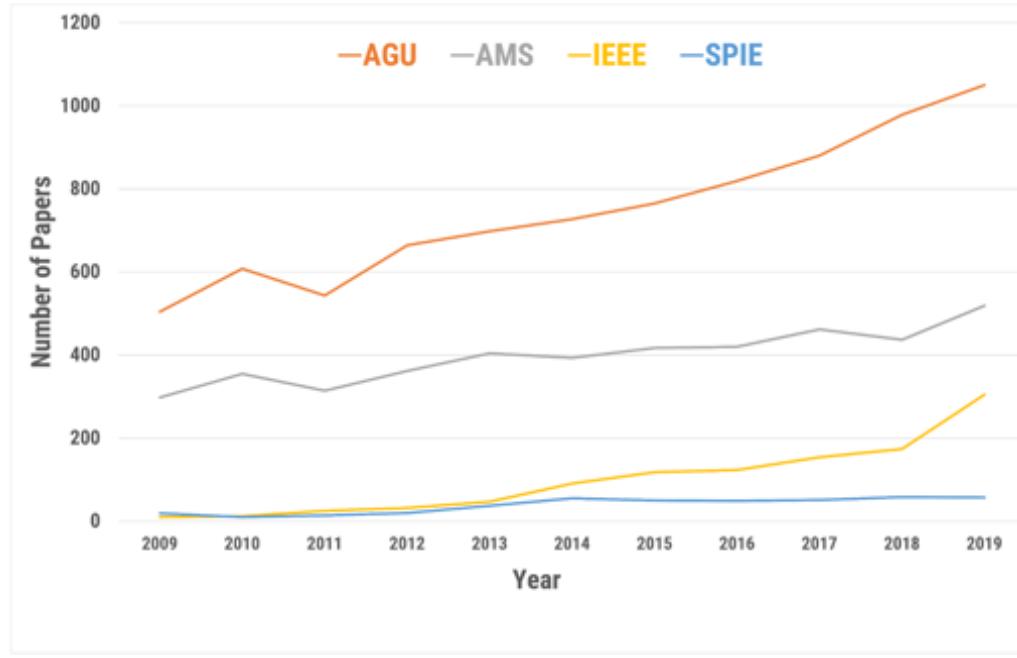
## AGU topic trend

Number of journal articles per keyword by year



2009

# Publication trend - ML in Earth sciences



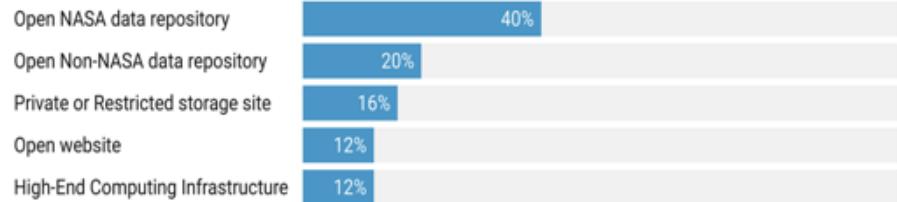
Rapid adoption of AI/ML by Earth science researchers

Virts et al. (2020)

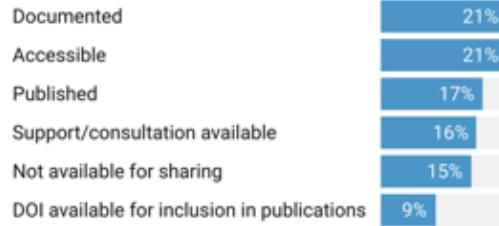
Maskey et al. (2020)

# NASA Science survey - AI and data

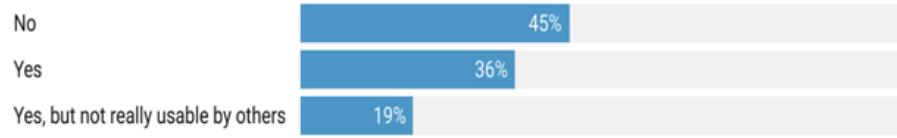
## Source of data used



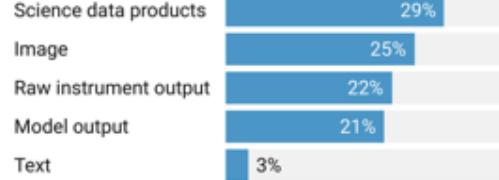
## How re-usable is your training data?



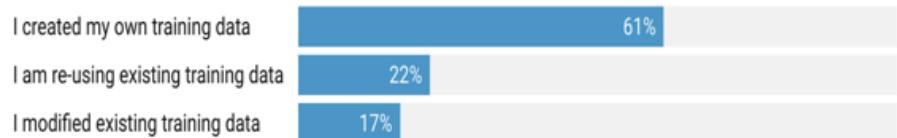
## Is there a catalog of training data for your use?



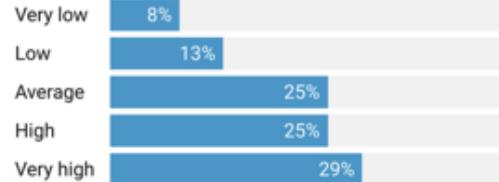
## What type of data do you use for AI?



## How did you construct training data?



## Amount of effort required to prepare data for AI?

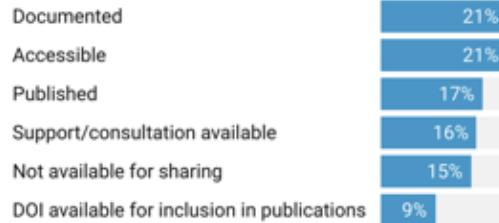


# NASA Science survey - AI and data

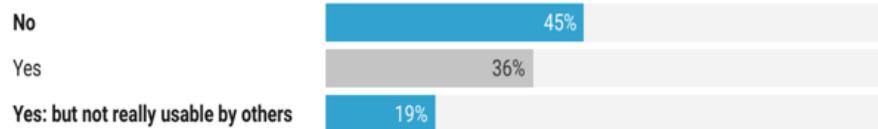
## Source of data used



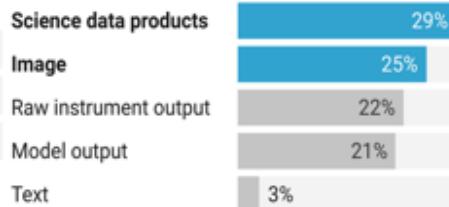
## How re-usable is your training data?



## Is there a catalog of training data for your use?



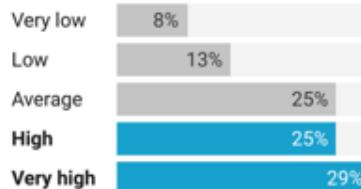
## What type of data do you use for AI?



## How did you construct training data?



## Amount of effort required to prepare data for AI?



# Maximizing Knowledge Discovery

# Why?

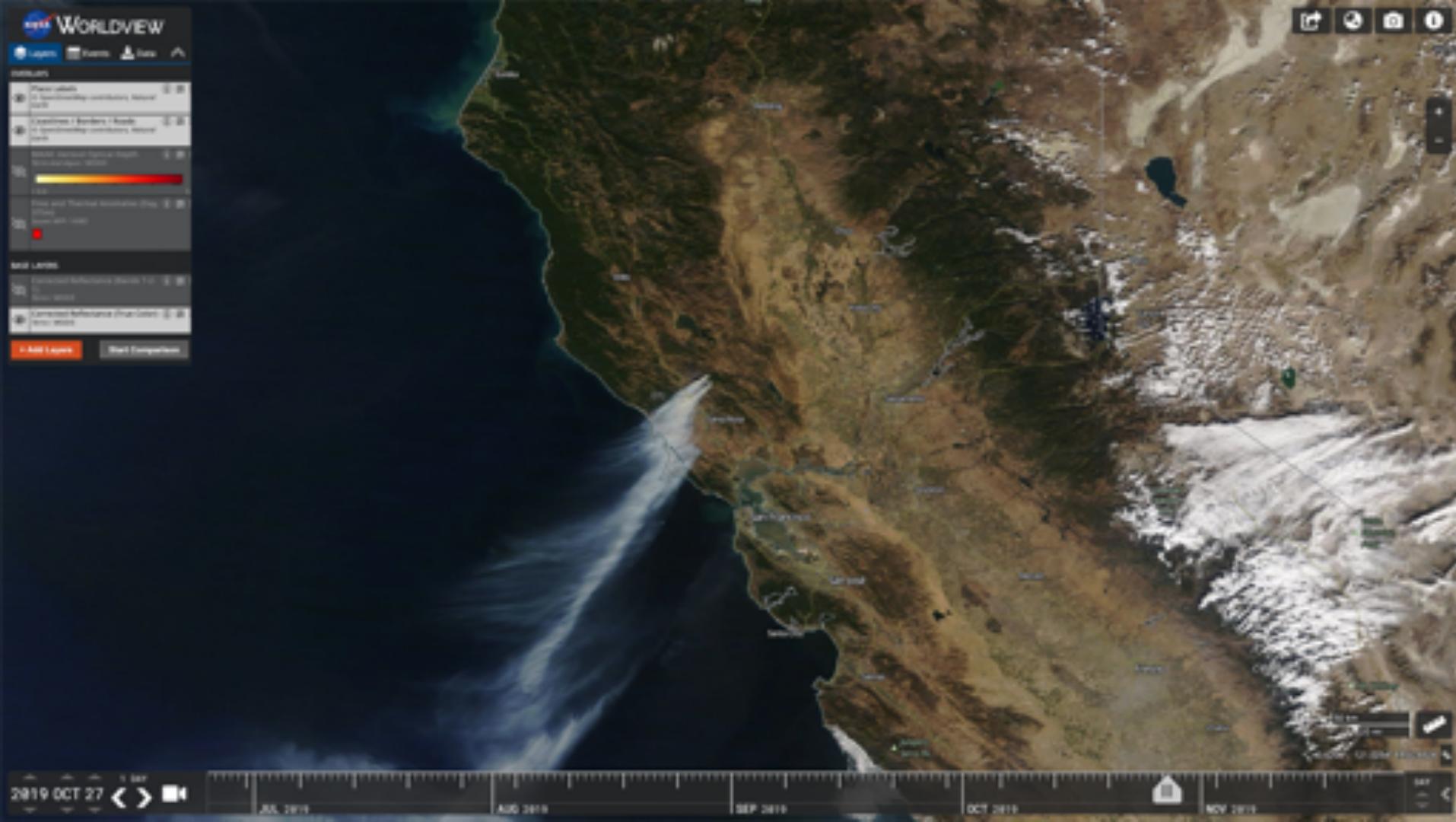
Increasing Earth science data archives require non-traditional approaches to data management

Data driven technologies (AI) to provide advanced search capabilities

Machine learning-based approach - provide automated detection of Earth science events from image archives

Catalog of events can provide a novel way to explore large archives of data

Discover and explore Earth science data archives around events using machine learning (ML) techniques



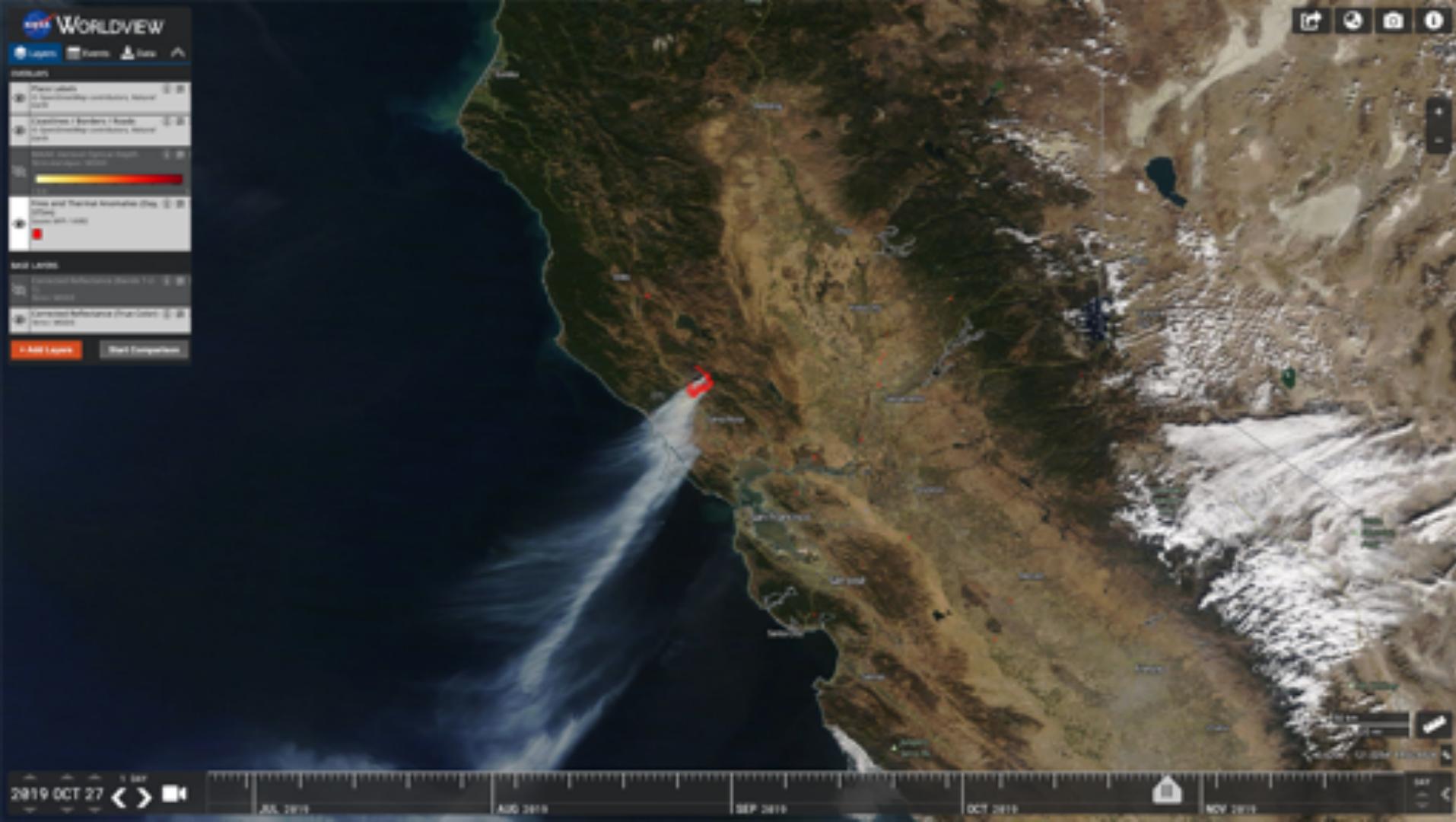
LAYERS

- Clouds and Cloud-Free Mask (Day) [✓]
- Clouds and Cloud-Free Mask (Night)
- Clouds and Thermal Anomalies (Day) [✓]
- Clouds and Thermal Anomalies (Night)
- Snow and Thermal Anomalies (Day) [✓]
- Snow and Thermal Anomalies (Night) [✓]
- Snow and Thermal Anomalies (Night)

BASE LAYERS

- Constrained Reflectance Model [✓]
- Cloud Mask
- Constrained Reflectance Prior Model
- World Cities

+ Add Layers Start Composition



LAYERS

- Placeholder (1)
- California - Georeferencing - constituency, National
- California - Borders - Admin
- California - Georeferencing - constituency, National

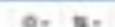
BASE LAYERS

- California Reference Grids (1)
- Clouds
- Compressed Reference Grids (1)

+ Add Layers Start Composition

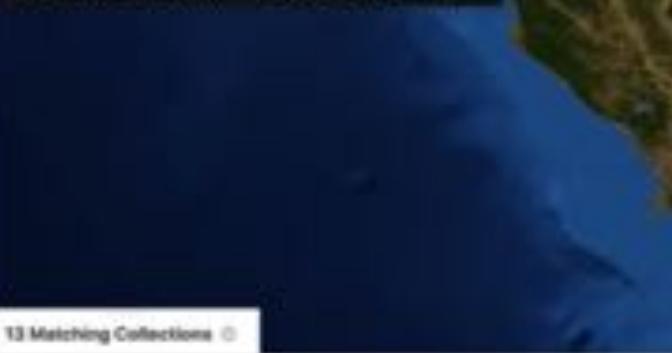
Earthdata  
Search

File



1. Point: 30.0199020000, -71.0562000000

2. Start: 2019-10-23 00:00:00      End: 2019-11-07 23:59:59



## 13 Matching Collections

Sort by: [Relevance](#)  Only include collections with granules  Include non-PCRSOS collections

Type: [Layer](#) Collection (13) selected by consumer and downloaded (0) file(s)



## MERRIT/Forest Thermal Anomalies/Fire 8-Day L3 Global 8km Grid 0000

A derivative of 00000-00-00 imagery + The Forest Watcher Resolution Imagery (ForestWatch) dataset contains Thermal anomalies and Fire 8 Day (MERRIT) data are generated at 1-kilometer (km) spatial resolution on a Level 3 product. The MERRIT product composite contains the maximum value of individual fire pixel sources detected during the eight days of acquisition. The Science Dataset (SDS) layers include the fire mask, user-quality indicators, maximum fire radiative power (MFRP), and the location of the fire event within the scene.

[View Details](#) [Download](#)



## MERRIT/Forest Thermal Anomalies/Fire Daily L3 Global 8km Grid 0000

A derivative of 00000-00-00 imagery + The Forest Watcher Resolution Imagery (ForestWatch) dataset contains Thermal anomalies and Fire Daily (MERRIT) data are generated every eight days at 1-kilometer (km) spatial resolution on a Level 3 product. MERRIT contains eight consecutive days of fire data concatenated, packaged into a single file. The Science Dataset (SDS) layers include the fire mask, user-quality indicators, maximum fire radiative power (MFRP), and the location of the fire event within the scene.

[View Details](#) [Download](#)



## MERRIT/Forest Thermal Anomalies/Fire 8-Day L3 Global 8km Grid 0000

A derivative of 00000-00-00 imagery + The Forest Watcher Resolution Imagery (ForestWatch) dataset contains Thermal anomalies and Fire 8 Day (MERRIT) data are generated at 1-kilometer (km) spatial resolution on a Level 3 product. The MERRIT product composite contains the maximum value of individual fire pixel sources detected during the eight days of acquisition. The Science Dataset (SDS) layers include the fire mask, user-quality indicators, maximum fire radiative power (MFRP), and the location of the fire event within the scene.

[View Details](#) [Download](#)



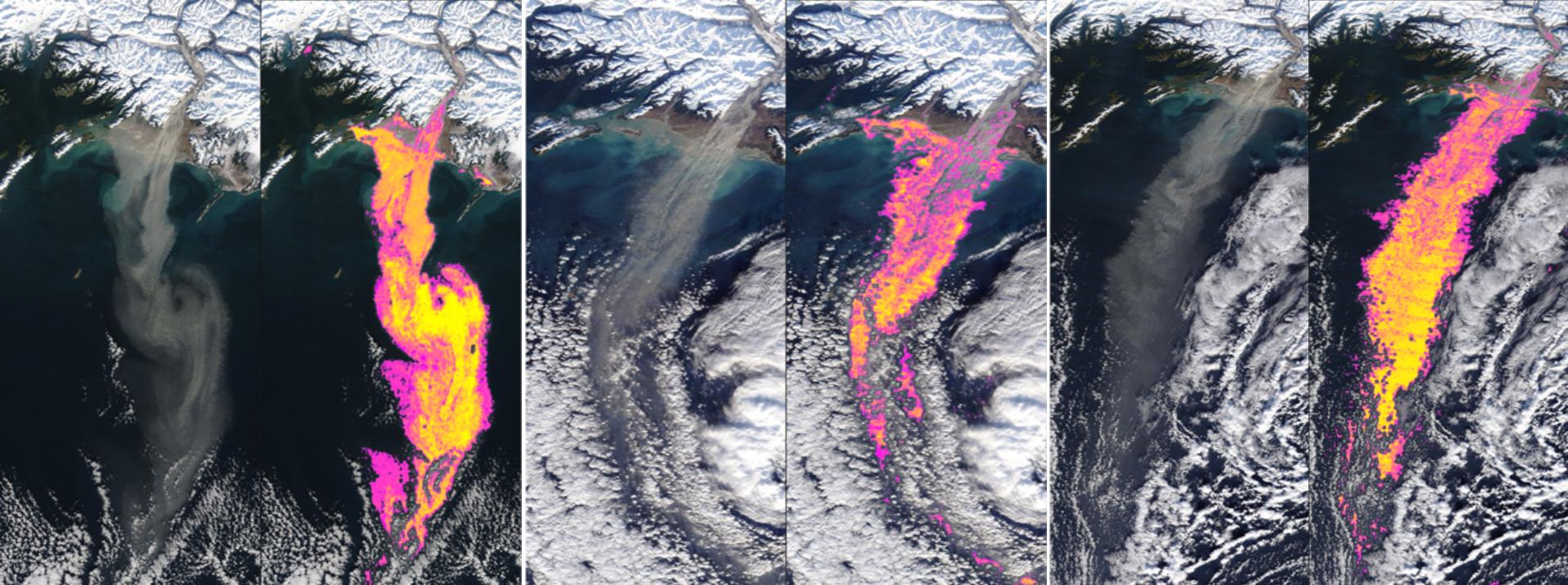
## MERRIT/Forest Thermal Anomalies/Fire 8-Day L3 Global 8km Grid 0000

A derivative of 00000-00-00 imagery + The Forest Watcher Resolution Imagery (ForestWatch) dataset contains Thermal anomalies and Fire 8 Day (MERRIT) data are generated at 1-kilometer (km) spatial resolution on a Level 3 product. The MERRIT product composite contains the maximum value of individual fire pixel sources detected during the eight days of acquisition. The Science Dataset (SDS) layers include the fire mask, user-quality indicators, maximum fire radiative power (MFRP), and the location of the fire event within the scene.

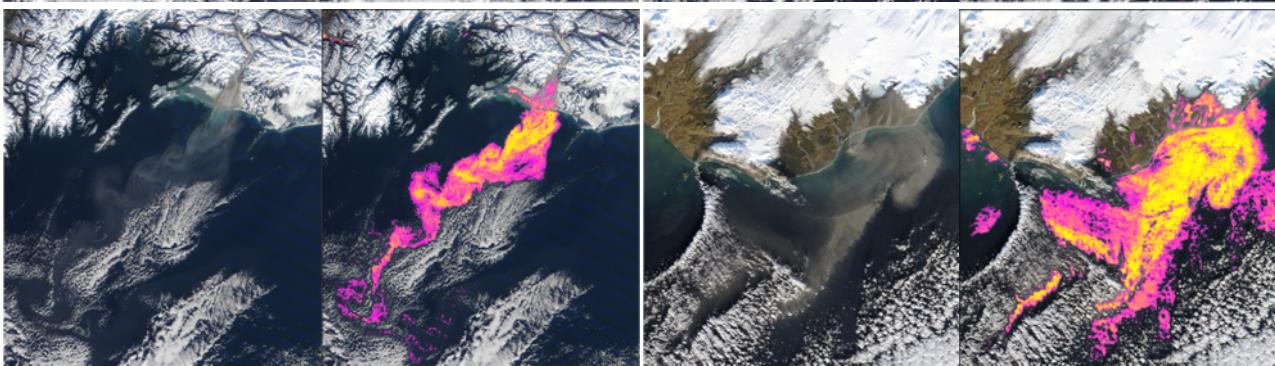
[View Details](#) [Download](#)



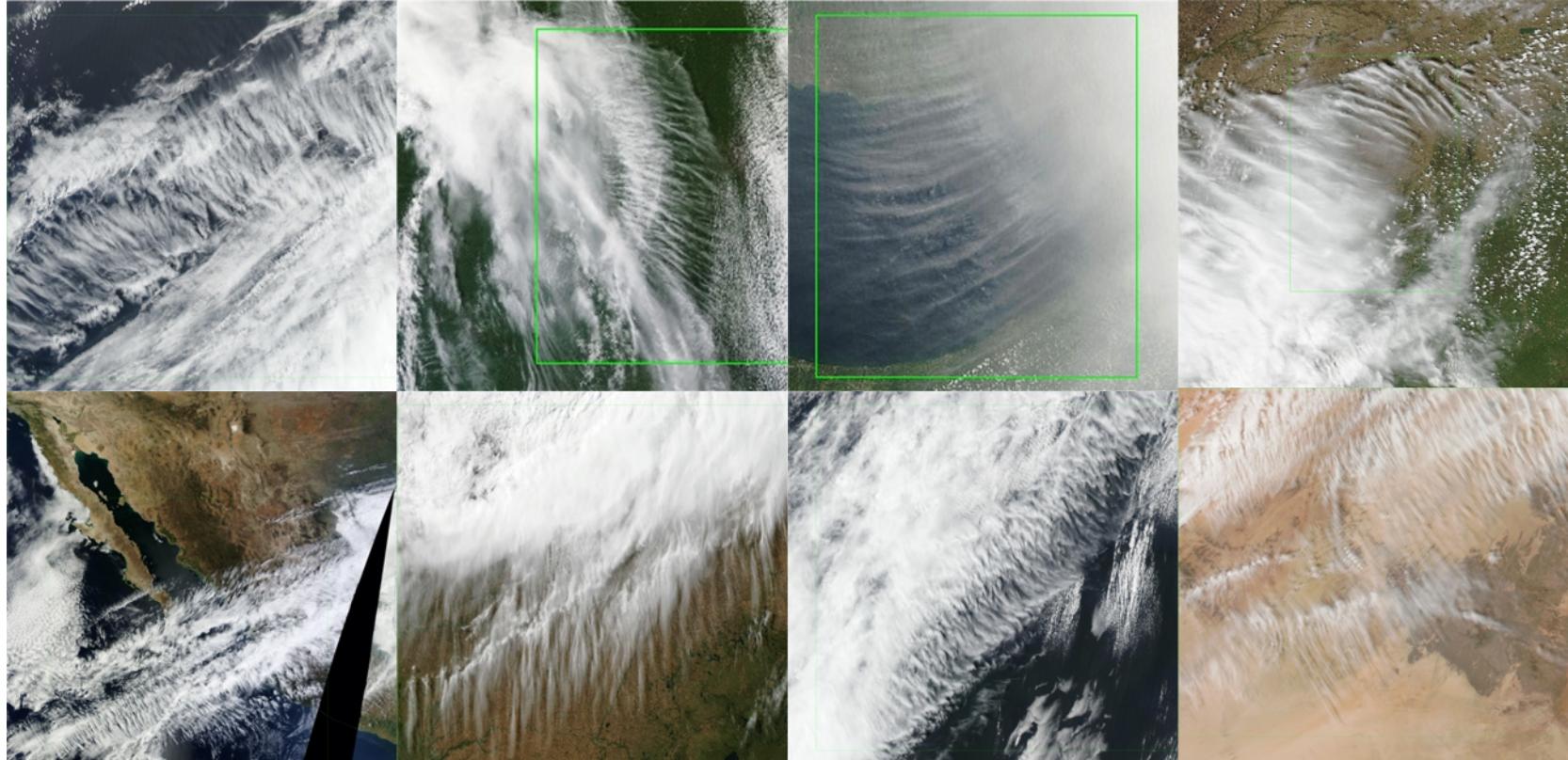




High latitude dust



# Transverse cirrus bands



Welcome to the  
**Phenomena  
Detection Portal**

We are using machine learning for real-time  
detection of Earth science phenomena:

Types  
03  
so far

Detections  
**98,627**  
and counting

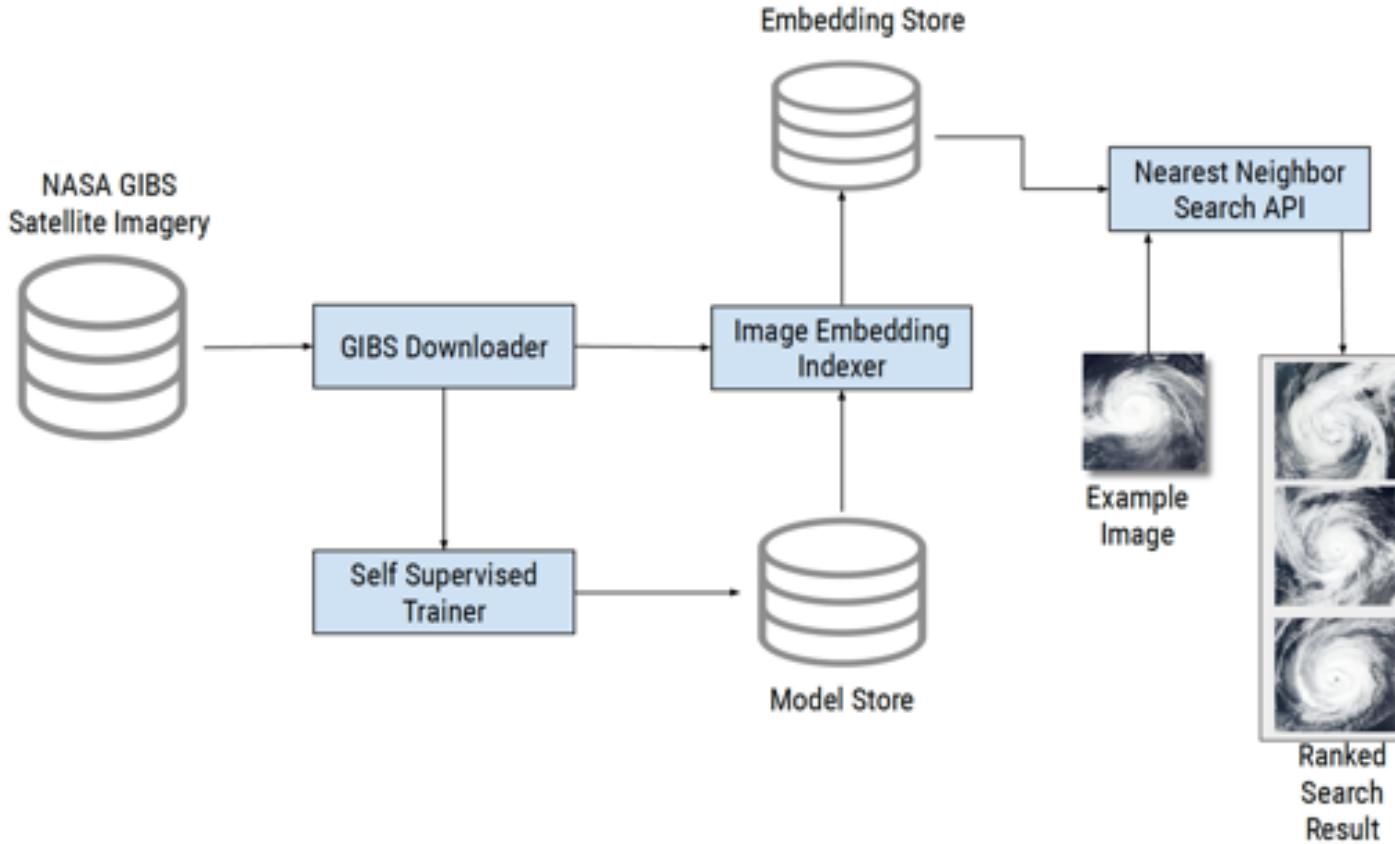
Confidence score  
**89.61%**  
on average

[Start exploring](#)

[Learn more](#)



# Search by Example



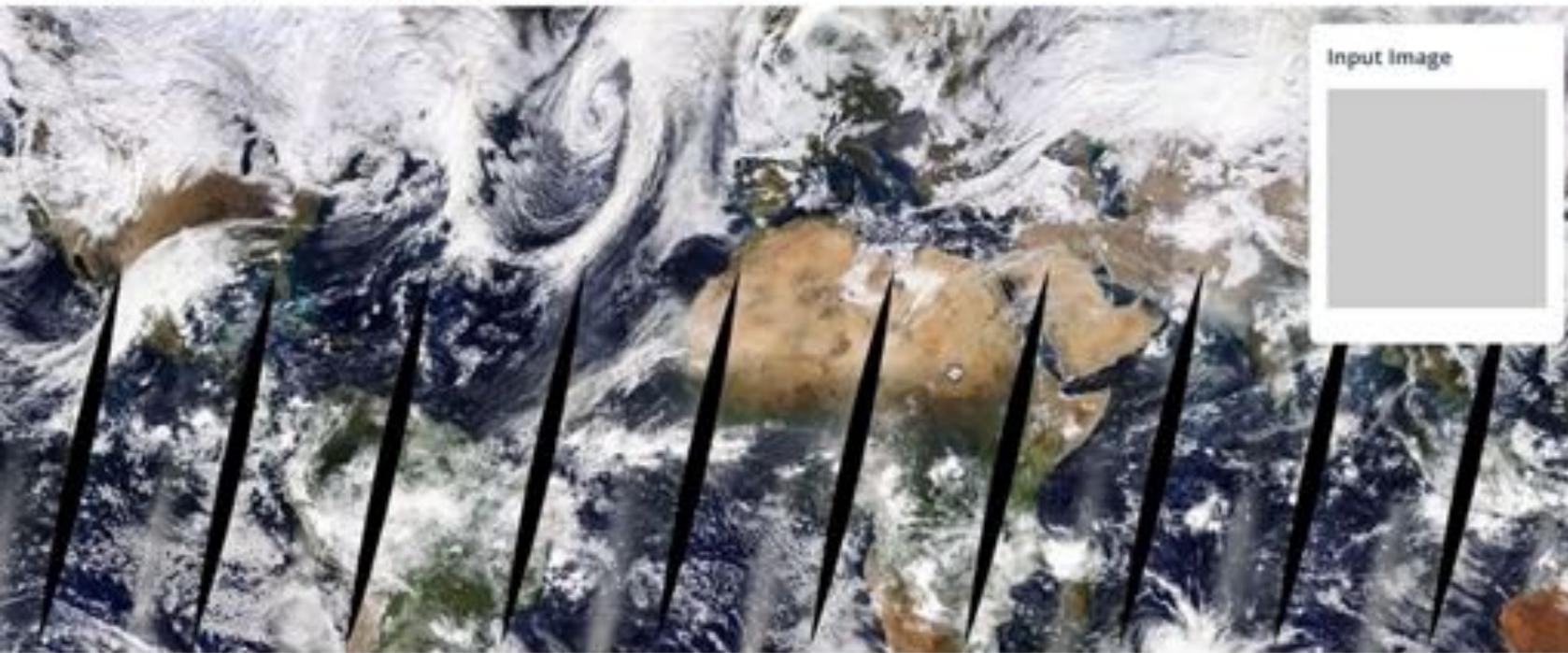
# Similarity Search Demo UI

CLICK THE MAP TO SELECT AN INPUT IMAGE - IMAGE DATE:

1/20/22

IMAGE NEIGHBORS: 1

+



Results



2010-07-11

Scaled  
SSL

# Augment data stewardship processes

# Why?

Assigning science keywords is currently a manual process, which is prone to human error and inconsistencies.

Metadata managed across a network of multiple data centers (i.e. keywords not assigned by a central entity)

Keywords may be assigned by non-subject matter experts (SMEs)

Improve metadata quality

Provide objective and consistent approach to keyword assignment



## Abstract

The LIS/OTD 2.5 Degree Low Resolution Annual Climatology Time Series (LRACTS) consists of gridded climatologies of total lightning flash rates seen by the spaceborne Optical Transient Detector (OTD) and Lightning Imaging Sensor (LIS). The long LIS (equatorward of about 38 degree) record makes the merged climatology most robust in the tropics and subtropics, while the high latitude data is entirely from OTD. The LRACTS dataset include annual flash rate time series data in MP4 format.

## DOI

10.5067/LIS/LIS-OTD/DATA306

## Science Keywords

EARTH SCIENCE > Atmosphere > Atmospheric Electricity > Lightning

EARTH SCIENCE > Atmosphere > Weather Events > Lightning

# Approach – build word embeddings

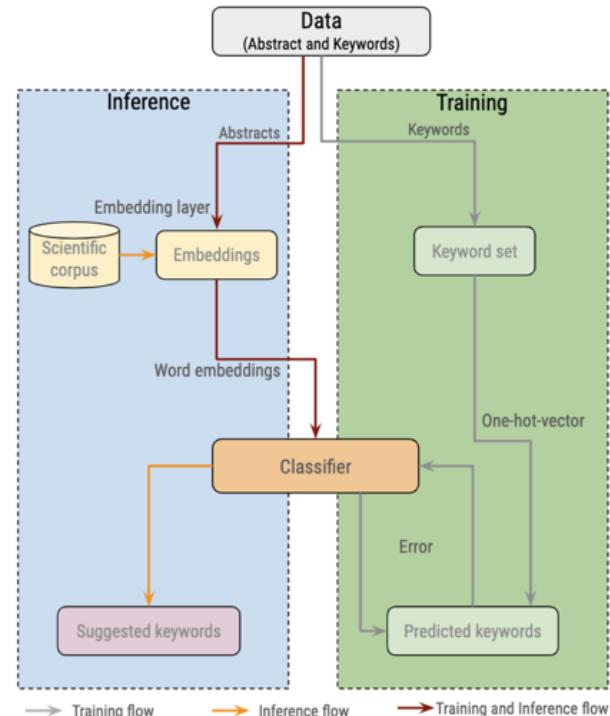
Journal Name	Date Published																					
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019				
Atmospheric Science Letters	5	21	34	27	27	42	39	59	64	34	80	69	67	69	65	67	58					
Earth and Space Science											1	24	28	25	42	36						
Earth's Future											13	26	24	52	58	60	56					
Eos, Transactions American Geophysical Union											46	37		1								
Geochemistry, Geophysics, Geosystems	89	297	296	373	195	148	298	574	164	293	337	242	196	154	305	567	64	29				
GeoHealth																	22	36	54			
Geophysical Research Letters	3,155	3,436	3,518	3,686	3,700	3,513	3,559	3,290	3,005	3,234	3,058	3,354	3,200	3,388	3,491	3,390	3,507	3,136				
Global Biogeochemical Cycles	132	136	324	326	75	66	94	66	83	68	96	76	72	76	75	74	23	21				
Journal of Advances in Modeling Earth Systems											6	3	9	20	25	49	88	66	513	138		
Journal of Geophysical Research											56		70									
Journal of Geophysical Research: Atmospheres	873	1,230	286	206	788	965	972	737	954	940	782	969	811	798	841	758	784	538				
Journal of Geophysical Research: Biogeosciences								23	29	140	393	312	146	185	130	92	109	158	138	45	24	
Journal of Geophysical Research: Earth-Surface								52	47	95	84	345	130	134	132	341	317	92	93	59	44	30
Journal of Geophysical Research: Oceans	253	497	375	314	317	341	438	321	382	513	418	338	345	300	325	388	492	284				
Journal of Geophysical Research: Planets	132	279	175	126	162	150	195	130	147	171	172	167	89	78	75	503	142	92				
Journal of Geophysical Research: Solid Earth	345	683	365	319	377	435	436	376	508	450	378	297	314	367	354	388	126	96				
Journal of Geophysical Research: Space Physics	439	561	496	525	475	447	589	533	756	654	694	496	503	541	592	542	466	446				
Meteorological Applications											7	47	69	76	76	46	61	67	76	5	2	
Paleceanography											65	109	96	62	67	61	86	47	68	59	50	
Paleceanography and Paleoclimatology																		24	26			
Quarterly Journal of the Royal Meteorological Society																						
RadioScience	100	132	346	314	94	122	93	508	59	136	6	185		203	168	157	178					
Reviews of Geophysics	9	23	12	15	8	54							12	12	14	23	16	22	36	57		
Space Weather								16	57	53	48	47	44	48	48	55	63	65	53	67	58	
Tectonics	55	85	79	88	58	73	66	60	71	47	76	58	21	83	99	119	45	58				
Water Resources Research	367	396	359	337	328	364	414	356	350	403	408	447	412	365	403	397	453	346				

88,410  
documents

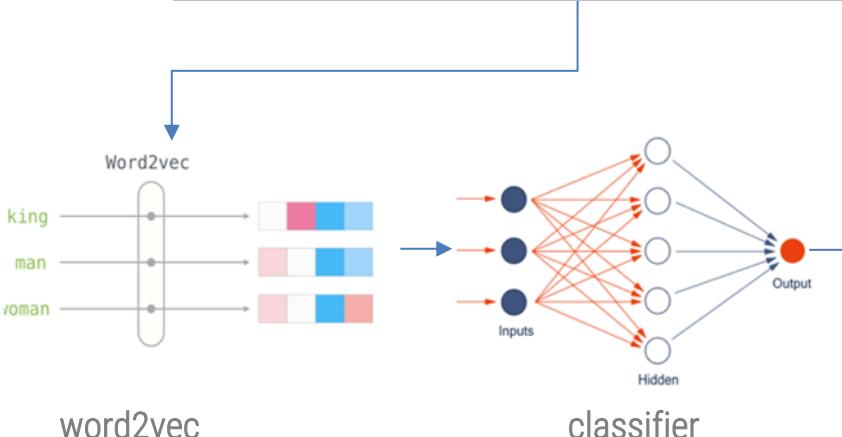
530 million  
words

5.5 million  
unique words

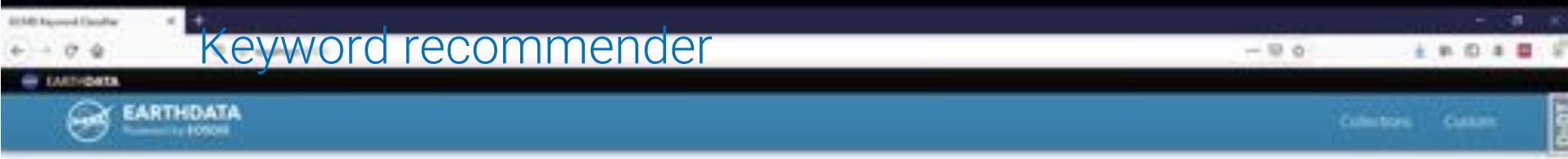
# Automated keyword assignment



Version 7.3 is the current version of the data set. Version 3.5 is no longer available and has been superseded by Version 7.3. This data set is currently provided by the OCO (Orbiting Carbon Observatory) Project. In expectation of the OCO-2 launch, the algorithm was developed by the Atmospheric CO<sub>2</sub> Observations from Space (ACOS) Task as a preparatory project, using GOSAT TANSO-FTS spectra. After the OCO-2 launch, "ACOS" data are still produced and improved, using approaches applied to the OCO-2 spectra. The "ACOS" data set contains Carbon Dioxide (CO<sub>2</sub>) column averaged dry air mole fraction for all soundings for which retrieval was attempted. These are the highest-level products made available by the OCO Project, using TANSO-FTS spectral radiances, and algorithm build version 7.3. The GOSAT team at JAXA produces GOSAT TANSO-FTS Level 1B

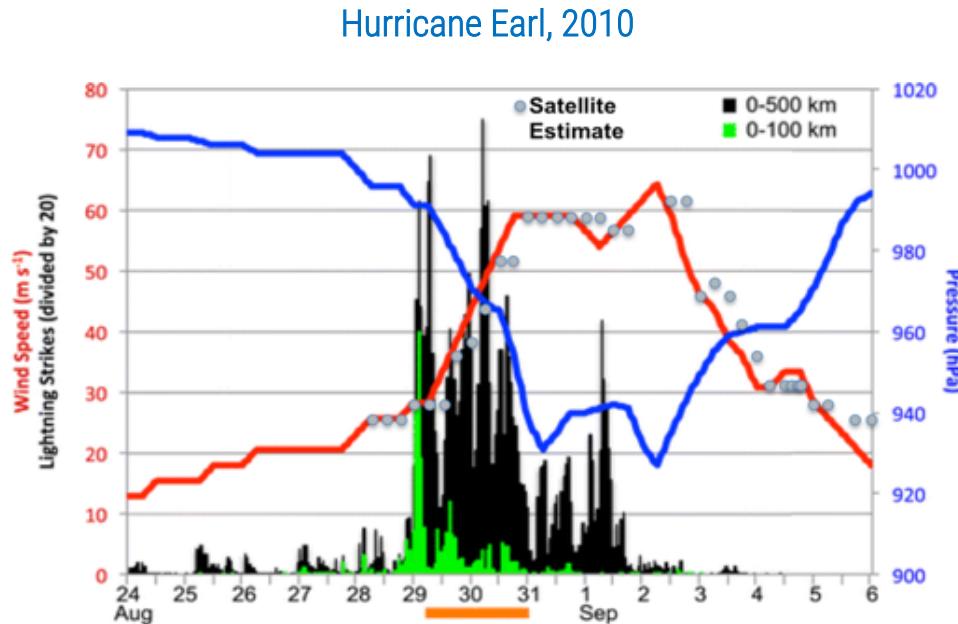


Predicted Keyword	Score
carbon dioxide	0.4513424
land use/land cover classification	0.3825603
terrain elevation	0.1924277
barometric altitude	0.18085223
carbon and hydrocarbon compounds	0.07634798



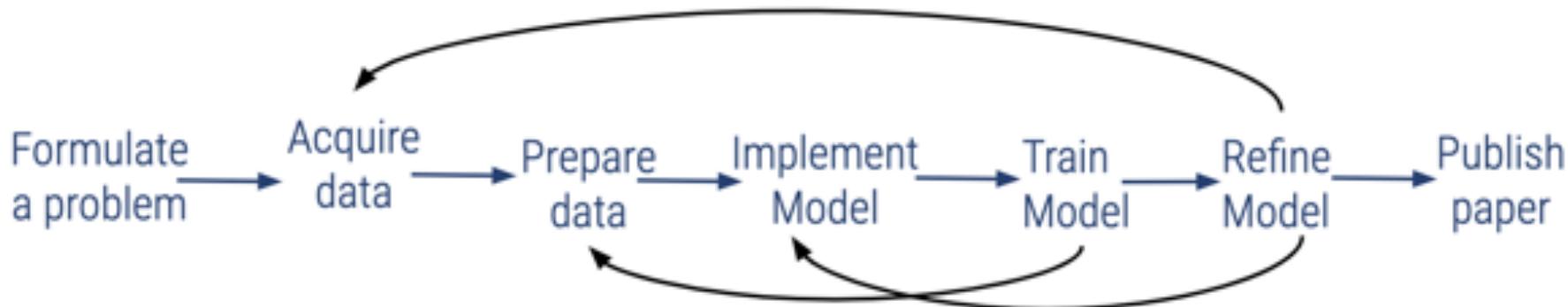
# Hurricane intensity estimation system

# AI and satellite imagery to estimate hurricane wind speed



Adapted from Stevenson et al. (2014). Time series of satellite-derived intensity estimates (circles) for Hurricane Earl (2010), added to best track intensities and lightning flash rate time series.

# ML in literature



# We have a model....now what?

Going extra mile

Interpretability + model inspection

Interpret prediction data – prediction output maybe just numbers

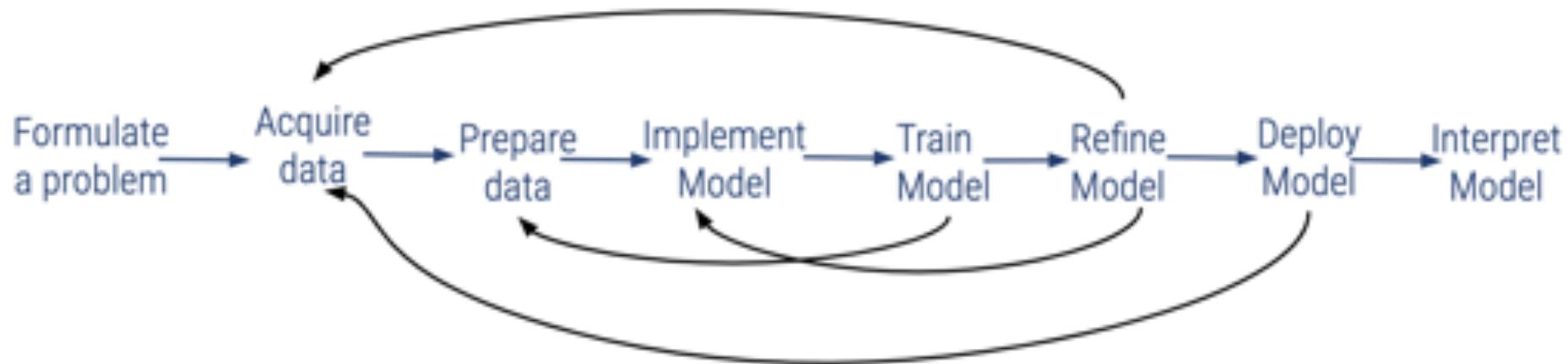
Questions:

Does the model confidence remain the same over time?

How do you maintain?

How do you complete the loop with new training data?

# ML lifecycle - iterative



# Deployment to production

## Performance requirements

Metrics and baselines with initial models

Monitor over time

## Back-testing

Model and software will change

Testing model changes on historical data

Run current production model to baseline performance

Run new models, competing for production

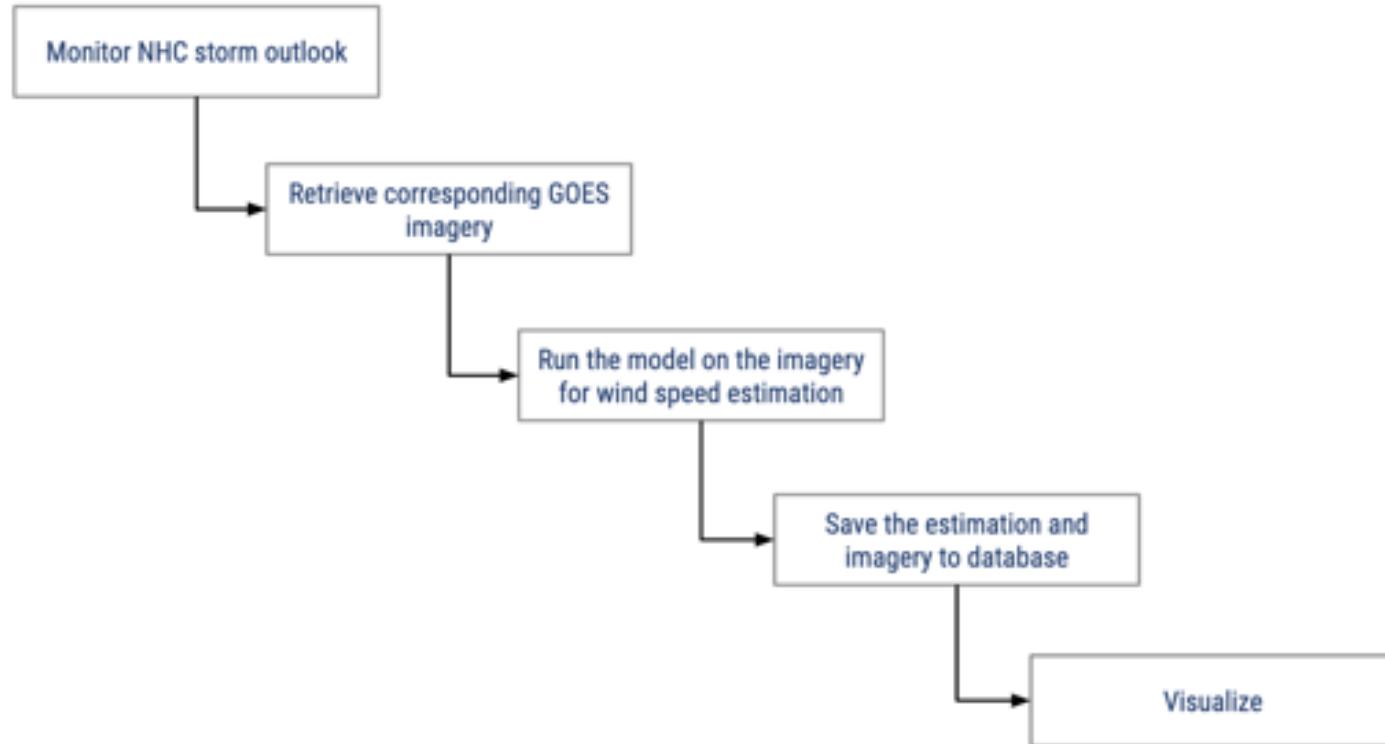
## Now-testing

Testing of production model on latest data

Can we get early warning that the model may be faltering?

- Content drift: training data exploited by model are subtly changing with time

# Workflow



## Deep Learning-based Hurricane Intensity Estimator

Applying machine learning to objectively estimate tropical cyclone intensity.

Explore

Read more

Start now

Using community to advance model development

# Data science competitions

- Benchmark datasets and challenge problems have played an important role in driving progress in AI
  - Enables rigorous performance comparison
- Foster the learning of best practices
- Stimulate the abilities in problem-solving
- Encourage creativity and group work
- Give learners the chance to interact with new platforms and algorithms
- Citizen science

# Data science competition

## “Wind-dependent Variables: Predict Wind Speeds of Tropical Storms”

- Leverage community to enhance solution to existing problem using open data
- Test whether high-quality datasets produces better models via open competition

- **Industry** partnership
- **733** participants
- **2756** entries

**Wind-dependent Variables: Predict Wind Speeds of Tropical Storms**  
HOSTED BY RADIANT EARTH FOUNDATION

Model shows that convolutional neural networks can capture key patterns in the satellite imagery of storms to estimate wind speed; we seek to improve the accuracy for operational applications.

**Task**

The goal of this challenge is to estimate the wind speeds of storms at different points in time using satellite imagery captured throughout a storm's life cycle across the entire range of storm sizes. Radiant Earth Foundation has worked with the Florida IMPACT team to assemble a data set of tropical imagery, which includes single-band satellite images at a long wave infrared frequency and corresponding wind speed annotations. Improving initial wind speed estimates from satellite imagery could mean significant improvements in short-term storm intensity forecasting, risk approximation models, and disaster readiness and response.

If the winning solution of this competition performs better than the existing model running on Hurricane Intensity Estimator, the model will be replaced with credit given to the winner.

Rank	Prize Amount
1st	\$3000 and \$6000 Amazon credit
2nd	\$4000 and \$8000 Amazon credit
3rd	\$2000 and \$4000 Amazon credit

This challenge is convened by our friends at Radiant Earth Foundation.

**Radiant Earth Foundation**  
EARTH IMAGERY FOR IMPACT

With generous support from:

**CONVENING SPONSOR**

 NASA  
NASA Earth Science Data Systems Program

**GOLD SPONSORS**

 aws  developmentSEED

**SILVER SPONSORS**

 azavea  Element 84  
AZURE CREDIT SPONSOR

 Microsoft

**TECHNICAL SUPPORTER**

 ESI

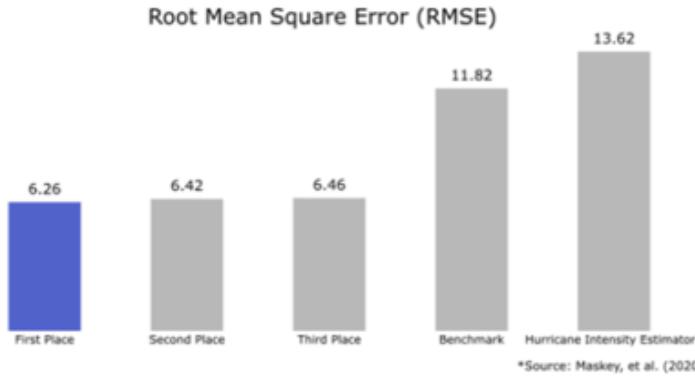
Banner image courtesty of the Centers for Disease Control and Prevention.

# Data science competition

## The Results

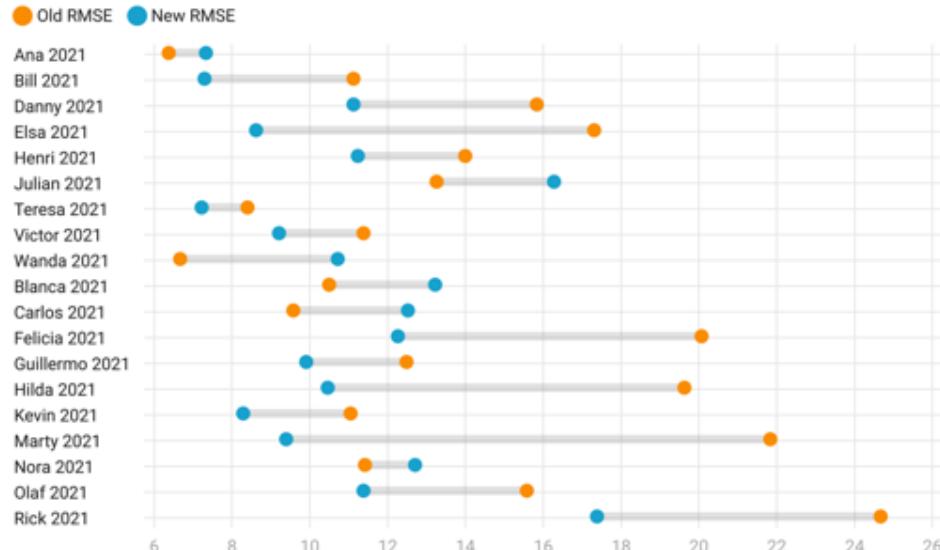
Over 700 participants stepped up to this important challenge, generating more than 2,700 entries.

**Each of the top three models achieved at least a 50% reduction in Root Mean Square Error (lower is better) as compared to the existing model!**

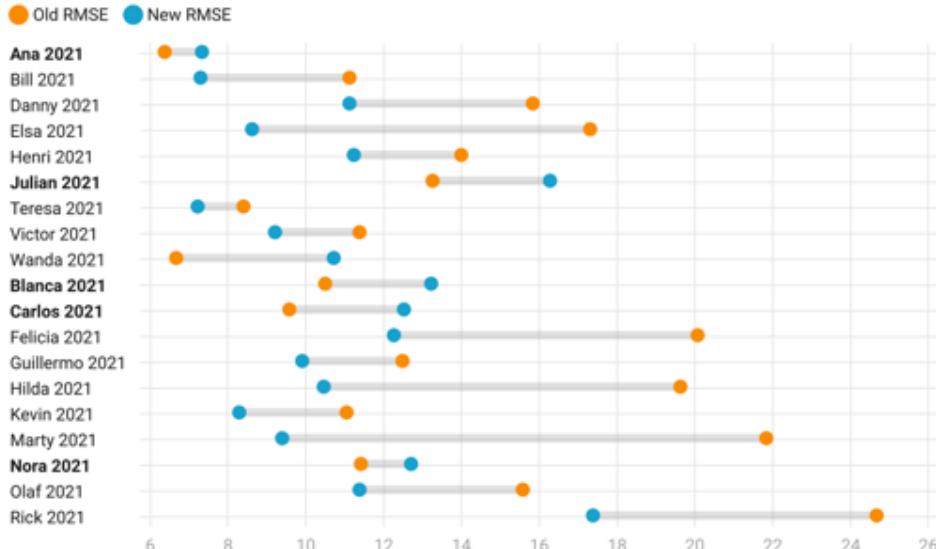


Winning solutions were able to take advantage of the relative timing of images in a storm sequence to produce targeted wind speed estimates based on temporal trends. As a result, these solutions can help to improve disaster readiness and response efforts around the world by equipping response teams with more accurate and timely wind speed measurements. All of the prize-winning solutions from this competition are linked below and made available for anyone to use and learn from.

# New Model – in depth analysis



# New Model – in depth analysis



# Flood extent detection

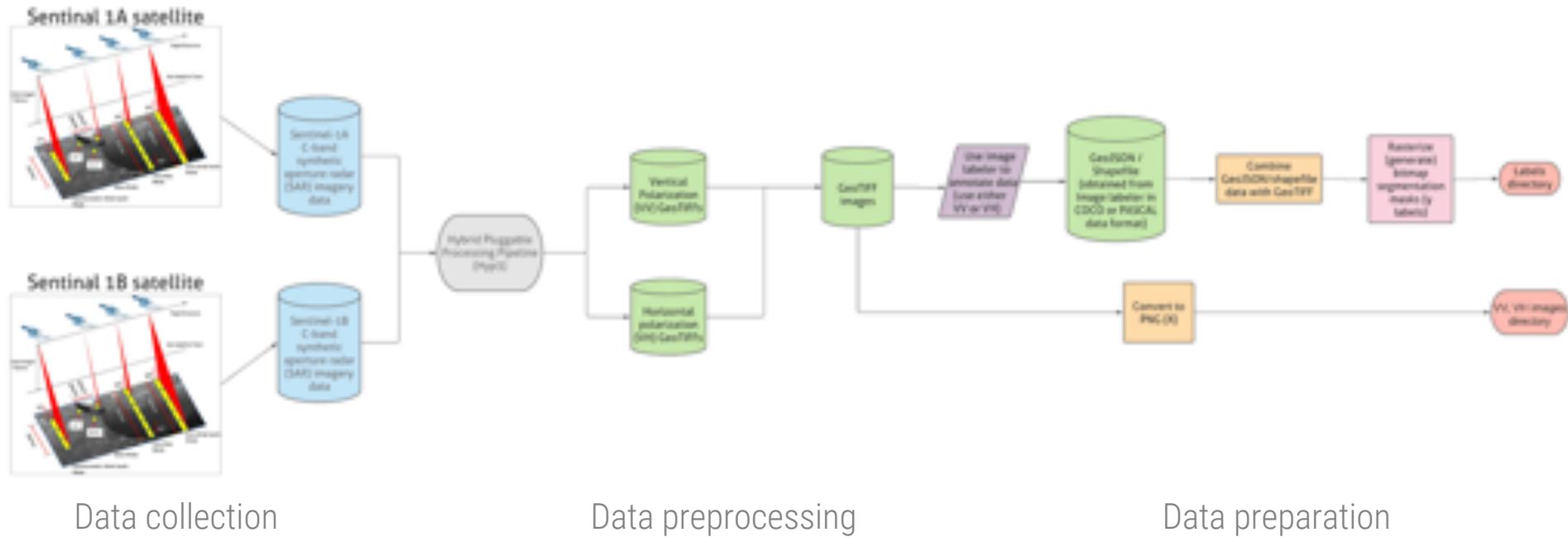
# Flood extent detection

- A major natural disaster
- Widespread damage – property, agriculture
- Displacement, insurance, long-term socio-economic consequences
- Causes:
  - Persistent rainfall
  - Severe storm
  - High-tides
  - Storm surge from cyclones

# Problem

- Detecting flood extent is difficult
- Monitoring extent of flood events in-situ – hazardous to operate in a disaster zone
- Potential solution:
  - Remote sensing in conjunction with ML has been used in the community to monitor these events
- Need:
  - Large amounts of clean and labeled data

# Data acquisition



# Data labeling

- 6 Atmospheric science/Earth science students
- 2 Domain scientists
- Training sessions
- Validation

ImageLabeler

WELCOME TO IMAGE LABELER

A web application  
to create and manage  
labeled Earth science images  
for machine learning.

CREATE A NEW SET

ABOUT THIS SITE

# Labeled data

~66k images (33k VV + 33k VH images including swath gap artifacts)

Native resolution : 5x20m

Train (24300):

- Nebraska (1741 sq. km.) (~43 %)
- North Alabama (13789 sq. km.) (~43%)
- Bangladesh (7150 sq. km.) (~13%)

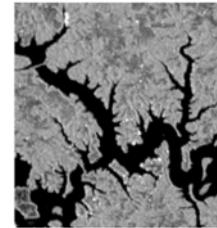
Validation (6500):

- Florence (7197 sq. km.)

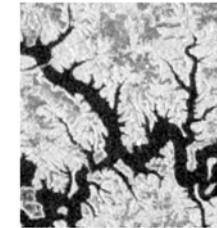
Test (1600):

- Red River North (6746 sq. km.)

VV

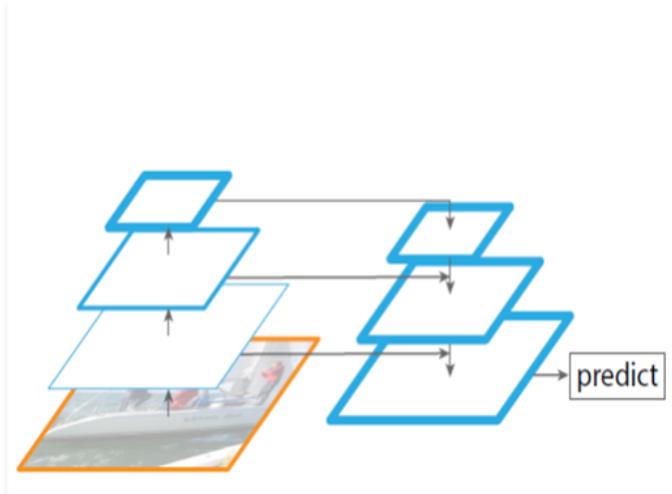


VH

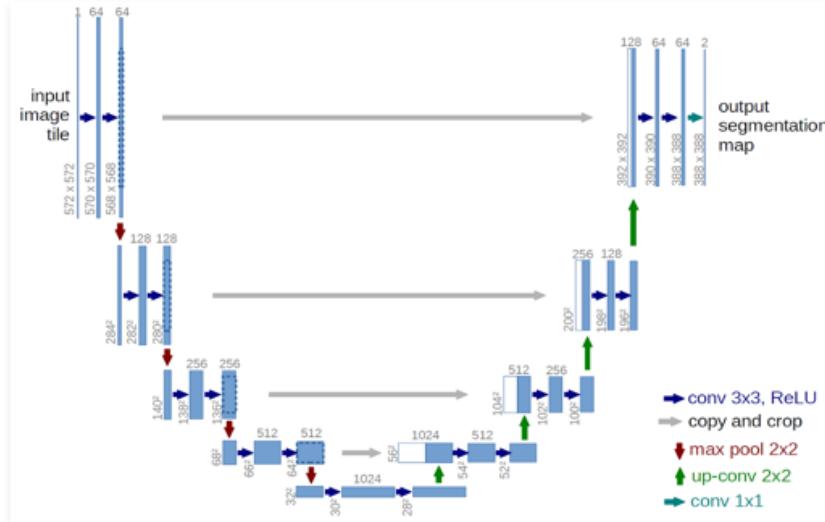


# Benchmark model development

Baseline models: Vanilla FPN and U-Net (ResNet 50 encoder). random selection, 4:1 Train-Validation split



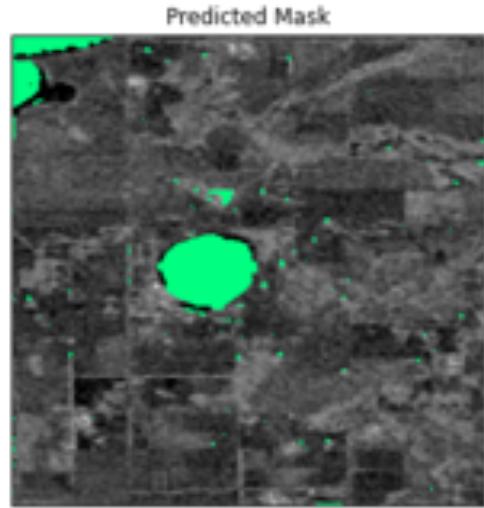
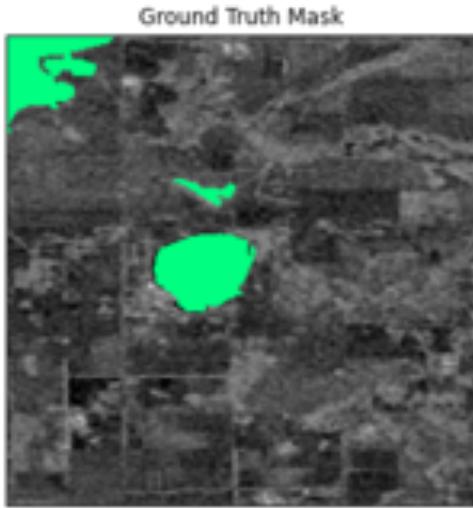
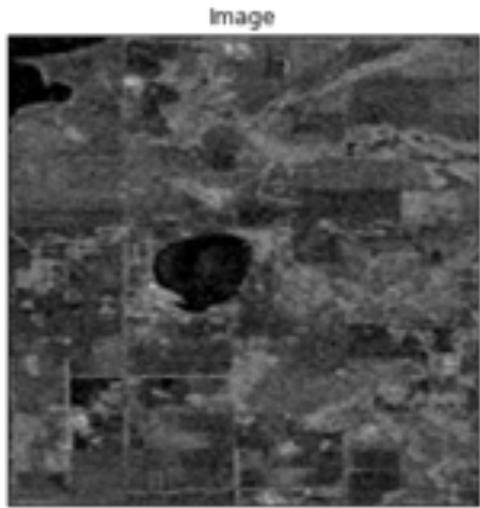
Feature Pyramid Network (Resnet 50 encoder)



U-Net (Resnet 50 encoder)

# Benchmark model development

## Visual Results



Sample segmentation

# Leveraging citizen science for optimal model

Finding optimal modal is an exhaustive task

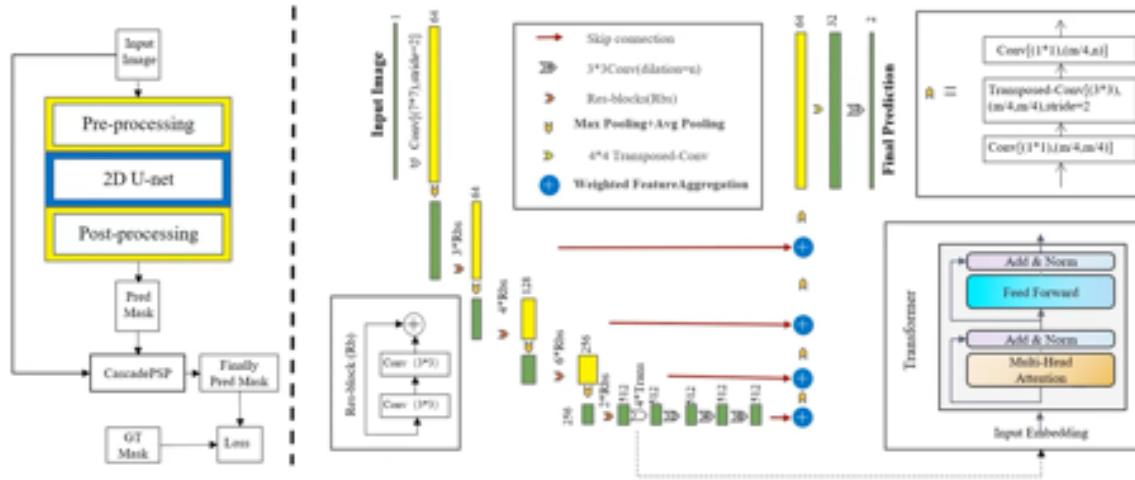
ML Competition in collaboration with IEEE

- 137 participants
- More than 200 submissions
- Codalab platform

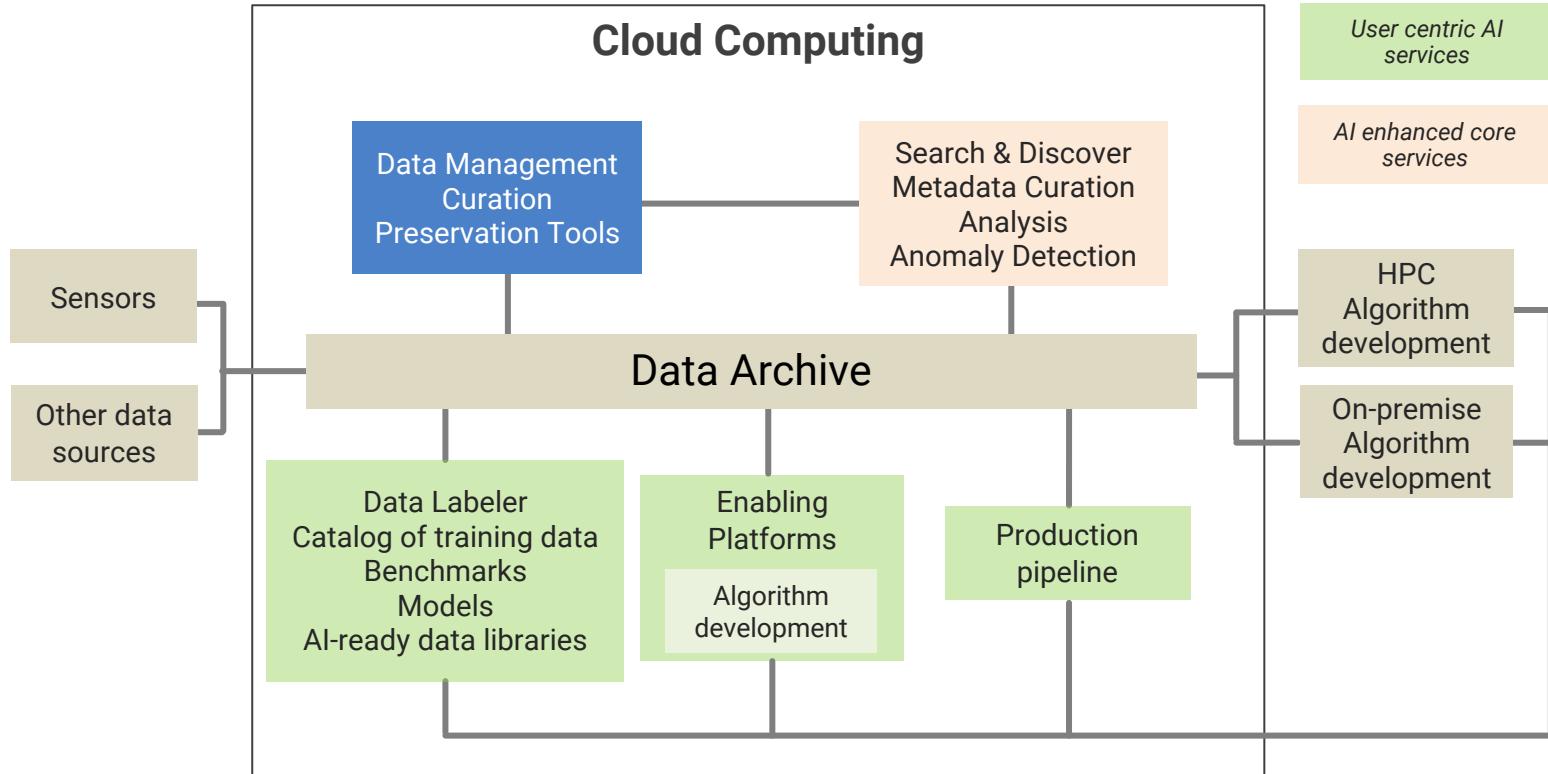
- **Phase 1 (Development):** Participants are provided with training data (which includes reference data) and validation data (without reference data until phase 1 concludes) to train and validate their algorithms. Participants can submit prediction results for the validation set to the codalab competition [website](#) to get feedback on the performance from April 15 to May 14, 2021. The performance of the best submission from each account will be displayed on the leaderboard.
- **Phase 2 (Test):** Participants receive the validation set reference data for model tuning and test data set (without the corresponding reference data) to generate predictions and submit their binary classification maps in numpy array format from May 15 to June 30, 2021. After evaluation of the results, three winners will be announced on July 1, 2021.

The screenshot shows the CodaLab competition interface for the ETCI 2021 Competition on Flood Detection. At the top, there's a navigation bar with 'Search Competitions', 'My Competitions', and 'Help'. The main title 'Competition' is centered above a thumbnail image of a satellite map showing flood areas. Below the title, it says 'ETCI 2021 Competition on Flood Detection' and 'Organized by Shubhankar - Current server time: Jan. 14, 2022, 2:35 a.m. UTC'. A timeline section shows 'First phase' (Development Phase 1, April 15, 2021, midnight UTC) and 'End' (Competition Ends, July 15, 2021, 11 p.m. UTC). Below this, a 'Learn the Details' sidebar has 'Overview' selected, along with 'Evaluation' and 'Terms and Conditions'. To the right, a large box welcomes participants to the competition, explaining the goal of detecting open water flood areas and the use of synthetic aperture radar (SAR) images. It also provides a link for more details: <https://nasa-impact.github.io/etc2021/>.

# Winning solutions



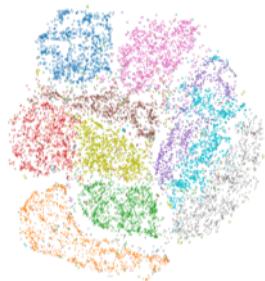
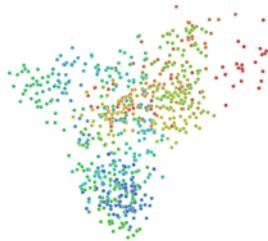
Team Arren, IOU: 0.7681



# Summary

## AI enhanced enterprise data systems

- AI approaches that can efficiently operate as a part of the core of large-scale systems
- Before AI can be widely used in critical enterprise data systems, we need new robust pipelines to systematically manage AI lifecycle
- A flexible architecture that allows software systems and AI algorithms to evolve to take advantage of emerging trends in hardware and software and rapid model deployment



Thank you.